The NIST 1999 Speaker Recognition Evaluation - An Overview

A. Martin and M. Przybocki

National Institute of Standards and Technology

alvin.martin@nist.gov mark.przybocki@nist.gov

Corresponding Author: Dr. Alvin Martin
NIST, Building 225
100 Bureau Dr Stop 8940
Gaithersburg, MD 20899-8940
USA

Phone: 1-301-975-3169
Fax: 1-301-670-0939

Abstract

This article summarizes the *1999 NIST Speaker Recognition Evaluation*. It discusses the overall research objectives, the three task definitions, the development and evaluation data sets, the specified performance measures and their manner of presentation, the overall quality of the results.

More than a dozen sites from the United States, Europe, and Asia participated in this evaluation. There were three primary tasks for which automatic systems could be designed: one-speaker detection, two-speaker detection, and speaker tracking. All three tasks were performed in the context of mu-law encoded conversational telephone speech. The one-speaker detection task used single channel data, while the other two tasks used summed two-channel data.

About 500 target speakers were specified, with two minutes of training speech data provided for each. Both multiple and single speaker test segments were selected from about 2000 conversations that were not used for training material. The duration of the multiple speaker test data was nominally one minute, while the duration of the single speaker test segments varied from near zero up to sixty seconds. For each task, systems had to make independent decisions for selected combinations of a test segment and a hypothesized target speaker. The data sets for each task were designed to be large enough to provide statistically meaningful results on test subsets of interest. Results were analyzed with respect to various conditions including, duration, pitch differences, and handset types.

Keywords

speaker recognition, speaker verification, speaker detection, speaker tracking, DET Curve, NIST evaluation

1 Introduction

The *1999 NIST Speaker Recognition Evaluation* was the latest in an ongoing series of yearly evaluations conducted by NIST. These evaluations aim to provide an important contribution to the direction of research efforts and the calibration of technical capabilities. They are intended to be of interest to all researchers working on the general problem of text independent speaker recognition. To this end the evaluations are designed to be simple, to focus on core technology issues, to be fully supported, and to be accessible.

This evaluation focused on speaker detection and tracking tasks in the context of conversational telephone speech. The evaluation was designed to foster research progress, with the objectives of:

1. exploring promising new ideas in speaker recognition,
2. developing advanced technology incorporating these ideas, and
3. measuring the performance of this technology

There were 16 participating sites in the 1999 evaluation, some working in collaborative partnerships. Included were four sites from the United States, one from India, and eleven from Europe. Six of the European sites or partnerships shared some resources and algorithms in a cooperative arrangement known as the ELISA Consortium [9]. Figure 1 summarizes the organizations that participated and the tasks for which they submitted system results.

Information on the NIST speaker recognition evaluations, including the official evaluation plans for past evaluations, is available on the NIST web-site [1].

2 The Tasks

The 1999 evaluation consisted of three separate tasks, two of which were new to the annual NIST speaker recognition evaluations. Each task consisted of several thousand trials. A *trial* consisted of a single hypothesized speaker and a specific test segment. The system was required to make an actual (true or false) decision on whether (or in the case of the speaker tracking task, where) the specified speaker was present in

the test segment. Along with each actual decision, systems were also required to provide for each trial a likelihood score indicating the degree of confidence in the decision. Higher scores indicated greater confidence in the presence of the speaker. A trial where the hypothesized speaker was present in the test segment (correct answer "true") is referred to as a *target trial*. Other trials (correct answer "false") are referred to as *impostor trials*.

2.1 One-Speaker Detection

This task is the basic speaker recognition task used in the previous NIST evaluations. The task is to determine whether a specified speaker is speaking in a given single channel segment of mu-law encoded telephone speech. The hypothesized speakers are always of the same sex as the segment speaker.

2.2 Two-Speaker Detection

This task is the same as the one-speaker detection task, except that the speech segments include both sides of a telephone call with two speakers present and the channels summed, rather than being limited to a single side and speaker. Note that the task is to determine whether the one specified speaker is present in the combined signal. The segment speakers may be of the same or opposite sex, but the hypothesized speaker is always of the same sex as at least one of the segment speakers.

2.3 Speaker Tracking

This task is to perform speaker detection as a function of time. Systems were required to identify the exact time intervals when the hypothesized speaker was speaking. For each identified time interval, an actual decision and a likelihood score were required. The task used a subset of the multiple speaker test data that was used for the two-speaker detection task.

3 Data

The data for the evaluation came from the Switchboard-2 Phase 3 Corpus collected by the Linguistic Data Consortium (LDC) [2]. This corpus consists of 2,728 five-minute conversations involving 640 speakers. The participating speakers were mainly college students in the southern United States and did not know one

another. They were recruited through campus flyers and local newspaper advertisements and paid a nominal fee for their participation. Each speaker was allowed to initiate and receive at most one call per day. Received calls were always on the same phone line, while initiated calls were required to all be made from distinct phone lines. Speakers were paired through an automaton that was called by the person initiating the call. A suggested topic was provided, but because subjects were not required to talk about the topic, college type chitchat dominated the conversations.

3.1 Training

Training data was provided for all hypothesized speakers of all trials. Such speakers are referred to as *target speakers*. The training data for each target speaker consisted of one minute of speech from each of two different conversations using the same phone number and thus, presumably, the same handset. Each one-minute segment consisted of consecutive turns of the speaker with areas of silence removed. This data was always selected from near the end of the conversation. (The end seemed a better, more conversational, choice than the beginning, which might contain more formal introductions.) The actual durations of the training segments were allowed to vary in the range of 55-65 seconds so as to include only whole turns wherever possible.

Training data was created for each speaker in the corpus for which there were two available conversations using the same phone number. This resulted in 230 male target speakers, and 309 female target speakers.

3.2 One-Speaker Detection

The one-speaker detection task used single speaker test segments. These test segments consisted of essentially the same data as in the two-speaker detection task described below, but with each of the two channels in these segments viewed as comprising a separate segment. As with the training segments, areas of silence were removed in the one-speaker test segments, and whole turns included to the extent possible. Thus, the length of these segments varied from close to zero up to a full minute. No more than one test segment was created from each conversation side, and no test segments came from conversations where training data was selected from either side. There were 3,420 one-speaker detection segments.

For each segment, eleven target speakers were specified for judgement as hypothesized speakers. One of these speakers was the actual speaker, provided that the actual speaker was a target speaker as was the case for more than 90% of the test segments. The other ten were randomly selected from among all target speakers of the same sex as the true speaker. Thus the total number or trials in this task was 37,620 with 3,157 of them being target trials and the remaining 34,463 being impostor trials. A target trial occurs when the hypothesized speaker is the speaker in the test segment and an impostor trial is when the hypothesized speaker is not.

3.3 Two-Speaker Detection

The two-speaker detection task used multiple speaker test segments. These test segments consisted of summed two-sided intervals with durations of 59 to 61 seconds selected from near the end of conversations, with whole turns included whenever possible. The duty cycle of target speakers varied from close to zero up to 100%. No more than one test segment was created from each conversation, and no test segments came from conversations where training data was selected from either side. There were 1723 two-speaker detection segments.

For each segment, twenty-two target speakers were specified for judgement as hypothesized speakers. These consisted of the two sets of eleven speakers used in corresponding single-sided segments in the one-speaker task. Note that two of the twenty-two were usually actual speakers in the conversation, and that the twenty-two speakers were either all of the same sex or consisted of eleven speakers of each sex, depending on the sexes of the two actual speakers of the test segment. The total number or trials in this task was 37,906 with 3,158 of them being target trials and the remaining 34,748 being impostor trials.

3.4 Speaker Tracking

The test segments consisted of a subset of 1000 of the test segments for the two-speaker detection task. Segments were chosen to include as many target speakers as possible and in fact 507 target speakers were represented in these segments. For each segment, the hypothesized speakers consisted of four of the twenty-two used in the two-speaker detection task. These included the two true speakers if they were target

speakers. If the two true speakers were of the same sex, all four hypothesized targets were of this sex; otherwise there were two hypothesized targets of each sex.

4 Performance Measure

Evaluation was performed separately for the three tasks. For each task the formal evaluation measure was a detection cost function, denoted C_{Det}, defined as a weighted sum of the miss and false alarm error probabilities as determined from a system's actual decisions:

$$(1)\ C_{Det} = (C_{Miss} * P_{Miss|Target} * P_{Target}) + (C_{FalseAlarm} * P_{FalseAlarm|NonTarget} * P_{NonTarget})$$

The required parameters in this function are the cost of a missed detection (C_{Miss}), the cost of a false alarm ($C_{FalseAlarm}$), the a priori probability of a target speaker (P_{Target}), and the a priori probability of a non-target speaker ($P_{NonTarget}$). In the 1999 evaluation the following parameter values were used:

$$(2)\ C_{Miss} = 10;\ C_{FalseAlarm} = 1;\ P_{Target} = 0.01;\ P_{NonTarget} = 1 - P_{Target} = 0.99$$

The relatively greater cost of a miss compared to a false alarm is probably realistic for many applications. The a priori probability of a target speaker was more arbitrary, and for the tracking task was probably not a realistic choice. Note that this specified a priori probability need not, and in fact did not, correspond to the actual percentage of target instances in the evaluation data.

For the detection tasks the decisions are discrete, and P_{Miss} and $P_{FalseAlarm}$ are defined by the counts of correct and incorrect decisions. For the tracking task, a reference answer key was determined by applying an energy-based automatic speech detector to each conversation side. P_{Miss} and $P_{FalseAlarm}$ were then computed as follows:

$$(3)\ P_{Miss} = \int_{\text{Target Speech}} \delta(D_t,F)dt \Big/ \int_{\text{Target Speech}} dt \qquad P_{FalseAlarm} = \int_{\text{Impostor Speech}} \delta(D_t,T)dt \Big/ \int_{\text{Impostor Speech}} dt$$

where D_t = the system output as a function of time, and

$\delta(x,y) = \{ 1 \text{ if } x=y, 0 \text{ otherwise}\}$

Systems were free to specify tracking decisions and scores down to the centisecond level. This was perhaps a greater level of granularity than needed, and resulted in some very large submission files.

The performance of systems on a given task can be shown by bar charts of the C_{Det} scores. In such bar charts (an example may be seen in Figure 3 in Section 5) we show the separate contributions of the false alarm rates and the missed detection rates to the total scores. We also use the decision likelihood scores to find all of the possible operating points (false alarm and missed detection rate pairs) of each system, and present these as DET (detection error tradeoff) curves, a variant of ROC curves using a normal deviate scale for each axis. (See [3] for a discussion of DET curves.) We note with an open circle "O" the point corresponding to the actual decisions, and with a diamond "◊" the point that would produce a minimum C_{Det} value. Examples appear in Section 5.

5 Results

For each of the three tasks a subset of the full evaluation test set corresponding to the test conditions of greatest interest as specified in the evaluation plan was defined. Figure 2 gives the details of these primary conditions and the numbers of target and impostor trials and speakers they encompassed. The primary conditions included that the handsets used in both training and test have electret microphones, and that different phone lines, and thus presumably different handsets, be used for training and test in the target trials. (See sections 6.3 and 6.4 below for a discussion of handset differences and microphone types.) Note that phone lines are always different for impostor trials. The speaker durations were required to be neither especially long nor especially short (in the 15-45 second range). For the two-speaker detection and speaker tracking trials the primary conditions had the additional restrictions that both handset microphones used in the test be electret and that the two speakers be of the same sex.

Figure 3 shows a bar graph of the detection cost function scores (C_{Det}) for the actual decisions by each system in the one-speaker detection task, using the primary condition trials. Clearly, there is a great deal of variation in system performance for this task. In accord with our understanding with the participants, we are not identifying the various systems in this plot or the DET plots presented hereafter, but for the identity of the best performing system for each task see the discussion below.

Figure 4 presents the corresponding DET curves for the one-speaker detection task. Since the DET plots generally contain more information of interest in terms of performance results and tradeoff, we hereafter concentrate on these. Figures 5 and 6 plot DET curves of the various systems' performance for the two-speaker detection and the speaker tracking tasks respectively, on the primary condition trials.

It may be observed from Figures 4 and 5 that the two-speaker detection task proved to be considerably harder than the one-speaker task. To some degree this is to be expected. A likely approach to the two-speaker detection task is to seek to separate the speech of the two speakers (this is relevant to the tracking task as well) and then apply one-speaker detection to each separated set of segments. The hypothesized speaker is declared present in the combined segment if and only if it is declared present in either of the two separated segment sets. Such an approach was used by at least some automatic systems [8], and also by a human attempting the task. If a one-speaker detection operating point has error probabilities (p1FA, p1M) then, based on these probabilities and an independence assumption, one might predict for the two-speaker task the operating point (p2FA, p2M) where

$$(4) \qquad p2FA = 1 - ((1 - p1FA) ** 2) = (2 * p1FA) - (p1FA ** 2)$$

$$p2M = p1M * (1 - p1FA) = p1M - (p1M * p1FA)$$

These ideas are explored in Figure 7. For two evaluation systems, we present DET curves which show, for the same underlying speech segments (the two-speaker primary condition segments), actual performance results for the one-speaker task, predicted performance results based on the above approach for the two-speaker task, and actual performance results for the two-speaker task. We see that this approach predicts part, but not the greater part, of the performance degradation in going from the one-speaker task to the two-speaker task, and other factors must also be involved. Some further information on the nature of the increased difficulty of the two-speaker detection task may be inferred from the discussion of duration in Section 6.

Figure 6 shows that speaker tracking is a very difficult task for current systems. Performance varied little across the systems participating and appears to be not far in excess of random. For this reason the upper limit on the probability scales in the figure is higher than that in the preceding figures. While part of the

difficulty is caused by inexactness in the automatically determined speech boundaries, analysis shows this not to be a major factor in the poor performance. Further research on how to approach this task, and how best to score performance on it, is clearly appropriate.

While performance differences among the best performing systems were not especially large and the evaluation was not intended primarily as a competition, it was desired to give formal recognition to the systems that did best on each task as determined by the official C_{Det} actual decision metric. Financing was not available this year for cash awards for the researchers involved. Instead, handsome plaques, as shown in Figure 8, were created and presented to the lead investigators of the best performing system for the primary condition for each of the three tasks. These winners were:

- Dragon Systems, Inc. - Fred Weber - one-speaker detection task
- MIT-Lincoln Laboratory - Douglas Reynolds, Bob Dunn, and Tom Quatieri - two-speaker detection task
- MIT- Lincoln Laboratory- Douglas Reynolds, Bob Dunn, and Tom Quatieri - speaker tracking task

6 Factors Affecting Performance

Having access to the raw submission results of all participants gives NIST the opportunity to analyze these results in order to explore various factors that may affect recognition performance. Over the years NIST has explored the significance of numerous data, speaker, and channel attributes on overall performance. Here we discuss several of these that appear to have been of importance in the 1999 evaluation.

6.1 Duration

Previous NIST evaluations included separate tests in one-speaker detection for segments of differing duration, namely 30, 10 and 3 seconds. These showed, as might be expected, that performance was significantly greater for 30-second segments than for 10-second segments, while performance on 10-second segments significantly exceeded that on 3-second segments. This year the one-speaker segments averaged 30 seconds in duration but varied over a continuous range up to one minute.

Figure 9 displays DET curves of one-speaker detection performance by ranges of segment duration for one system. The results shown are representative of those for all of the systems in the evaluation. They indicate that, at least for systems of the type used in this evaluation, performance is significantly lower for segments of shorter than 15 seconds in duration, but that for segments longer that 15 seconds duration does not greatly affect the performance level. This is consistent with the finding of previous years, but indicates that the duration effect seen then is limited, and that once some minimum duration, apparently in the 10-15 second range, is available, the amount of test speech ceases to be a major factor in performance.

Figure 10 presents a somewhat similar plot for the two-speaker detection task for one system. For this task all segments have a duration of about one minute. The individual plots therefore each include all impostor tests but are limited to target tests where the duration of speech by the true speaker are in the specified ranges. Here it is seen that duration does matter for performance, at least up to the two highest ranges presented, involving target speaker durations in excess of 35 seconds. This held for all participating systems. For the two highest duration ranges the performance differences were small and not consistent in direction among the several participating systems. This suggests that target speaker duration matters at least up to the point where it is significantly in excess of the speech duration of the other speaker in the test segment.

6.2 Pitch

Pitch would appear to be an important factor in speaker recognition, but attempts to specifically include it in algorithms have had only limited success [4]. NIST has investigated ways in which average speaker pitch affects performance in each of the last several years. Results have not been consistent on how performance is affected by limiting consideration within each sex to speakers of particularly high or low pitch [5, 6]. We have found, as might be expected, that limiting impostor trials to instances where the impostor's average pitch is close to that of the hypothesized target (in the training data), while including all target trials, degrades performance. But, perhaps surprisingly, a bigger effect is observed when target trials are restricted to those with the largest pitch differences between the training and test segments, while all impostor trials are included.

Figure 11 gives an example of a typical system in the 1999 evaluation. For each speaker, the average pitch of the training data and of each test segment was estimated. The plot shows for the one-speaker detection task, a curve of all primary condition tests, and curves limited to target trials where the log pitch difference between the test segment and the target speaker training data are in the high and low 25% of all such differences. Large pitch differences in target trials may correspond to instances where the speaker had a cold or was feeling particularly emotional during either the training or test conversation. Note how large the performance differences are. For example, at a 10% miss rate, the false alarm rate is around 4% when all trials are included. When target trials are limited to the 25% that are closest in pitch, the false alarm rate is less than 1%; when limited to the 25% furthest in pitch, the false alarm rate exceeds 10%.

6.3 Handset Differences

The variation in telephone handsets is a major factor affecting the performance of speaker recognition using telephone speech. For the Switchboard-2 Corpus specific handset information is not provided, but telephone line information (i.e. the phone number) is. We can generally assume that if two calls are from different lines the handsets are different; if they are from the same line the handsets are probably the same.

We have chosen to concentrate in this evaluation on target trials where the training and test segment lines are different. This is certainly the harder problem. Moreover using same line calls is in a way unfairly easy, since for impostor trials the training and test segment handsets are always different as, with rare exceptions, speakers do not share handsets. Thus using same line target trials could be viewed as handset recognition rather than speaker recognition.

Figure 12 shows, for one system, the large performance difference in this evaluation between using same line and different line target trials. The impostor trials are the same in both cases.

6.4 Handset Type

Most standard telephone handset microphones are of either the carbon-button or electret type. We have seen in recent evaluation that the handset types (i.e., the microphone types) used, both in the training and the test segments, can greatly influence recognition performance.

Handset type information for this evaluation was determined using the MIT-Lincoln Lab automatic handset labeler [7]. This is a software package that uses the input telephone speech signal from one channel to assign a likelihood that the signal is from a carbon-button handset as opposed to an electret handset. This likelihood was converted into a hard decision (carbon or electret). This hard decision was made available to the systems processing the training and test segments. The decisions made were certainly less than perfect, as occasionally conversations from the same phone number were assigned opposite type, but are believed to be generally accurate.

Figure 13 provides some information on the distribution of handset types in the conversations used in the evaluation. The main point to note is that the received calls generally from home phones, overwhelmingly involve electret type handsets, while the initiated calls often made from public phones, are split between electret and carbon-button type. There is also evidence that conversation sides involving female speakers are more likely to be declared to be of electret type. This may indicate a slight bias in the automatic type detection algorithm. It may also help to explain the slightly better overall performance of most systems on male speakers compared to that on female speakers.

Figure 14 shows the variation in performance for different combinations of training and test segment handset types for a typical system. All target trials here are different line. There is a considerable advantage to having matching handset types, and particularly to having electret type handset microphones in both the training and test segments.

7 Future Plans

The 2000 NIST Speaker Recognition Evaluation is being planned. It is expected to reuse selected data from the various parts of the existing Switchboard corpora. It may also use a Switchboard-type corpus of cellular telephone data now being collected. It was suggested at the last workshop that some non-English data be included. We are checking for the possible availability of suitable such data.

We expect the three tasks of the 1999 evaluation to reappear in the 2000 evaluation. Some consideration is being given to variations of the speaker tracking task.

To learn more about future evaluation plans or to arrange to participate in the next evaluation, please visit the NIST web-site [1] or contact Alvin Martin [10].

REFERENCES

1. http://www.nist.gov/speech/spkrinfo.htm
2. Linguistic Data Consortium, University of Pennsylvania, 3615 Market Street, Suite 200, Philadelphia, PA 19104-2608. http://morph.ldc.upenn.edu.
3. Martin, A., et al. The DET curve assessment of detection task performance. *Proceedings EuroSpeeech* Vol. 4 (1997), 1895-1898.
4. Doddington G., et al. "NIST Speaker Recognition Evaluation - Overview, Methodology, System, Results, Perspective -", to appear in *Speech Communication*.
5. Przybocki, M. and Martin, A., NIST Speaker Recognition Evaluation-1997, *RLA2C,* Avignon, April 1998, pp. 120-123.
6. Przybocki, M. and Martin, A., RLA2C presentation, Avignon, April 1998, NIST Speaker Recognition Evaluations: Review of the 1997 & 1998 Evaluations, http://www.nist.gov/speech/rla2c_pres/index.htm.
7. Quatieri, T., Reynolds, D., O'Leary, G., Magnitude-Only Estimation of Handset Nonlinearity with Application to Speaker Recognition, *Proceedings ICASSP* (1998), 745-748.
8. Dunn, Robert B., Reynolds, D., Quarteri, T., Approaches to Speaker Detection and Tracking in Multi-Speaker Audio, in DSP, Vol. 10, n. 1, January 2000.
9. The ELISA Consortium, The ELISA Systems for the NIST'99 Evaluation in Speaker Detection and Tracking, in DSP, Vol. 10, n. 1, January 2000.
10. 100 Bureau Drive Stop 8940, Gaithersburg, MD 20899-8940, alvin.martin@nist.gov.

8 FIGURE CAPTIONS

Figure 1

Figure 1: **Participants**. Listed are the 16 sites that participated in the *1999 NIST Speaker Recognition Evaluation*. Represented are four sites from the United States, one from India, and eleven from Europe. Six groups submitted results for the one-speaker detection task only. One group submitted results for the both the one-speaker detection task and the speaker tracking task, while five groups submitted results for all three tasks. There were two collaborative partnerships across sites, and six of the European groups (identified by "•"), shared some resources and algorithms in a cooperative arrangement known as the ELISA Consortium.

Figure 2

Figure 2: Primary Conditions. The *primary condition* for each of the three evaluation tasks was defined to include the trials reflecting the test conditions of greatest interest for research on the core problems of speaker recognition as specified in the evaluation plan. Trials that satisfied these conditions formed the basic set on which NIST reported the official results.

Figure 3

Figure 3: Actual Decision Costs for One-Speaker Detection. This bar graph shows the contributions of each error type to the total detection cost for each system. The trials used were those satisfying the one-speaker detection primary condition.

Figure 4

Figure 4: DET Curves for One-Speaker Detection. This plot shows DET curves, encompassing the full range of operating points, for each system. For each curve, the actual decision operating point is marked with a diamond, while the minimum detection cost operating point is marked with an open circle.

Figure 5

Figure 5: DET Curves for Two-Speaker Detection. This plot shows DET curves for the five participating systems on the primary condition trials.

Figure 6

Figure 6: DET Curves for Speaker Tracking. This plot show DET curves for the six participating systems on the primary condition trials. The ranges included on each axis have been expanded from those in the other plots because of the generally low performance level on this task.

Figure 7

Figure 7: One-Speaker vs. Two-Speaker Detection. Performance is shown for each of two systems on the detection tasks using the same underlying primary condition data. Both the actual two-speaker performance and that predicted from the one-speaker performance are shown for each system.

Figure 8

Figure 8: Official Winners. Shown are the plaques that were presented to the system developers of the official winning systems for each of the three evaluation tasks. These winners were the systems with the lowest actual decision detection cost for each task on primary condition trials.

Figure 9

Figure 9: Effect of Segment Duration for One-Speaker Detection. Performance is shown by duration ranges for segments satisfying the one-speaker primary conditions other than segment duration. It is seen that performance is considerably poorer for segments shorter than 15 segments, but that for longer segments duration does not affect performance.

Figure 10

Figure 10: Effect of Segment Duration for Two-Speaker Detection. Performance by duration ranges is shown, as in Figure 9, for two-speaker detection primary condition segments. Here duration matters until the segment length exceeds 35 seconds.

Figure 11

Figure 11: Effect of Target Pitch Differences. Shown for one system is the primary condition performance and performance when target trials are restricted to those with the 25% closest and furthest average pitch differences between the training and test data. The effect of such differences on performance is quite large.

Figure 12

Figure 12: Effect of Same/Different Lines for Training/Test. Performance for a typical system is far higher when target trials are restricted to those involving the same phone line (and presumably handset) in training as in test as opposed to those involving different lines.

Figure 13

Figure 13: Handset Type Distribution in Corpus. Received calls generally involved electret handsets, but initiated calls are divided between handset types. Presumably this is because the initiators often used pay phones. Females apparently used fewer carbon button handsets than males. This may indicate a bias in the handset type determination algorithm.

Figure 14

Figure 14: Performance as a Function of Training/Test Handset. Performance for one system on different number tests for each combination of training and test handset types. Performance improves when the types match, and is superior for electret handsets.

Figure 1:

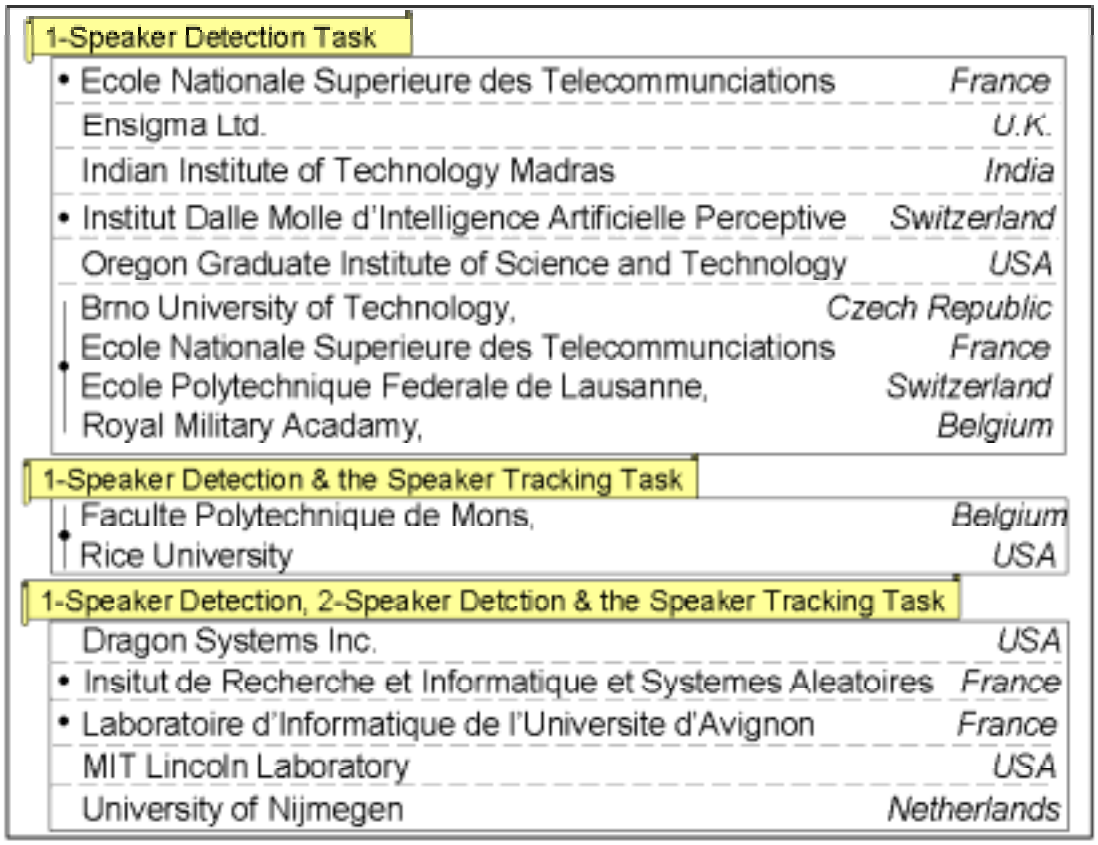
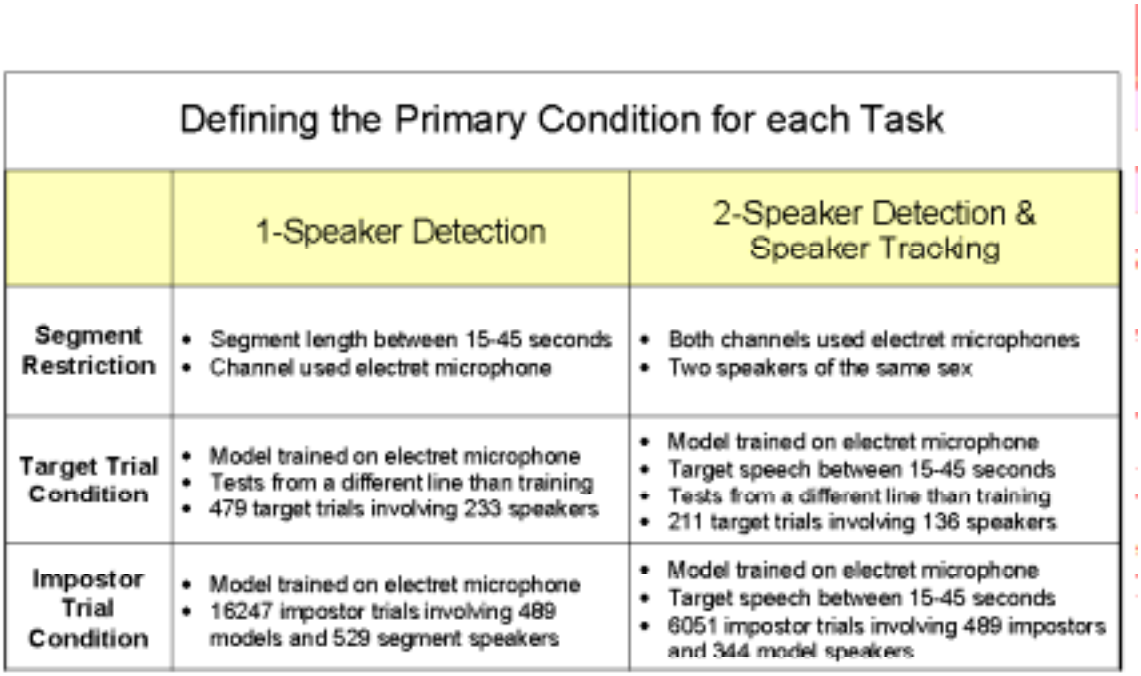

1-Speaker Detection Task
• Ecole Nationale Superieure des Telecommunciations France
Ensigma Ltd. U.K.
Indian Institute of Technology Madras India
• Institut Dalle Molle d'Intelligence Artificielle Perceptive Switzerland
Oregon Graduate Institute of Science and Technology USA
Brno University of Technology, Czech Republic
Ecole Nationale Superieure des Telecommunciations France
Ecole Polytechnique Federale de Lausanne, Switzerland
Royal Military Acadamy, Belgium
1-Speaker Detection & the Speaker Tracking Task
Faculte Polytechnique de Mons, Belgium
Rice University USA
1-Speaker Detection, 2-Speaker Detction & the Speaker Tracking Task
Dragon Systems Inc. USA
• Insitut de Recherche et Informatique et Systemes Aleatoires France
• Laboratoire d'Informatique de l'Universite d'Avignon France
MIT Lincoln Laboratory USA
University of Nijmegen Netherlands

Figure 2:



Defining the Primary Condition for each Task

	1-Speaker Detection	2-Speaker Detection & Speaker Tracking
Segment Restriction	• Segment length between 15-45 seconds • Channel used electret microphone	• Both channels used electret microphones • Two speakers of the same sex
Target Trial Condition	• Model trained on electret microphone • Tests from a different line than training • 479 target trials involving 233 speakers	• Model trained on electret microphone • Target speech between 15-45 seconds • Tests from a different line than training • 211 target trials involving 136 speakers
Impostor Trial Condition	• Model trained on electret microphone • 16247 impostor trials involving 489 models and 529 segment speakers	• Model trained on electret microphone • Target speech between 15-45 seconds • 6051 impostor trials involving 489 impostors and 344 model speakers

Figure 3:

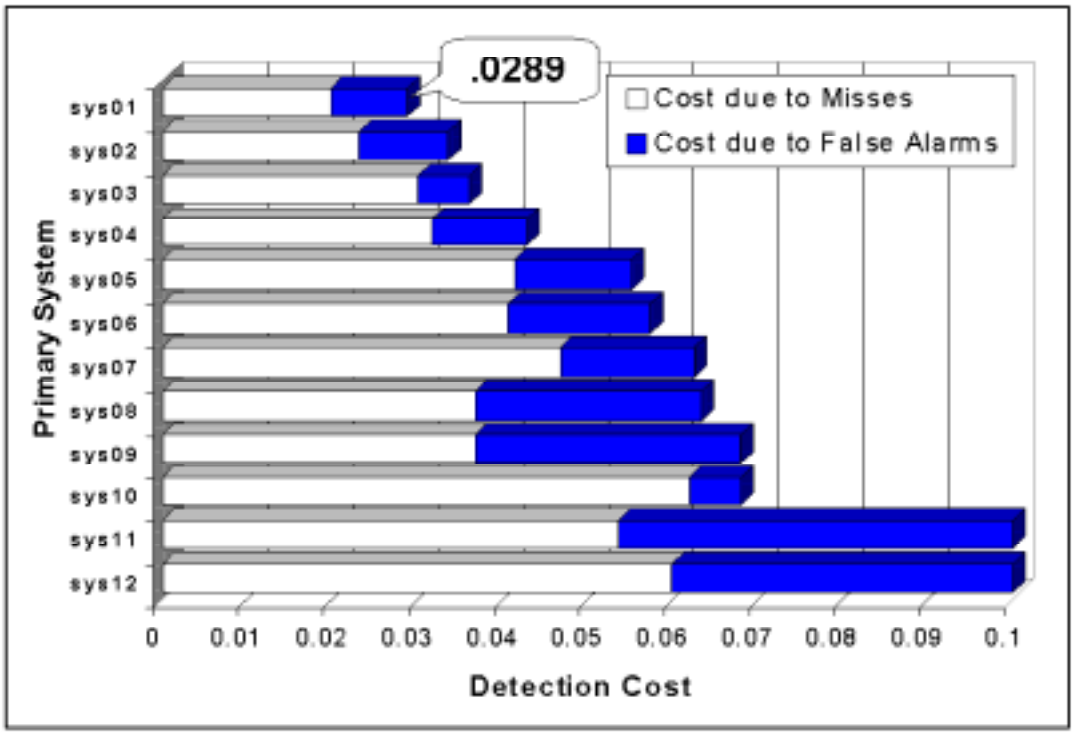

.0289
Cost due to Misses
Cost due to False Alarms
Primary System
sys01
sys02
sys03
sys04
sys05
sys06
sys07
sys08
sys09
sys10
sys11
sys12
0
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
0.09
0.1
Detection Cost

Figure 4

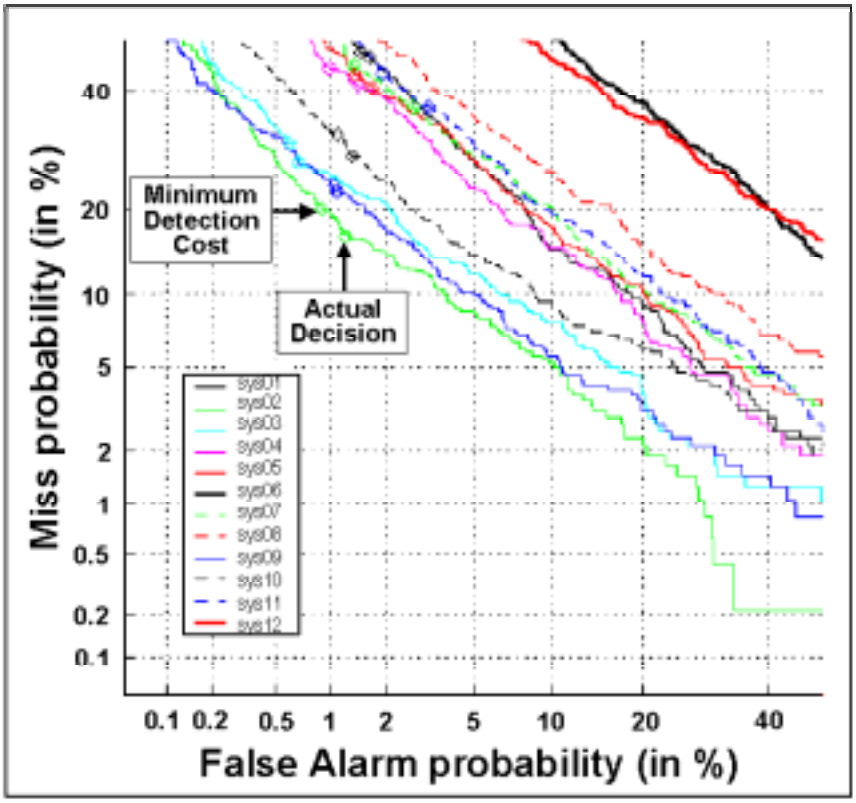

Miss probability (in %)
40
20
10
5
2
1
0.5
0.2
0.1
Minimum Detection Cost
Actual Decision
sys01
sys02
sys03
sys04
sys05
sys06
sys07
sys08
sys09
sys10
sys11
sys12
0.1 0.2 0.5 1 2 5 10 20 40
False Alarm probability (in %)

Figure 5

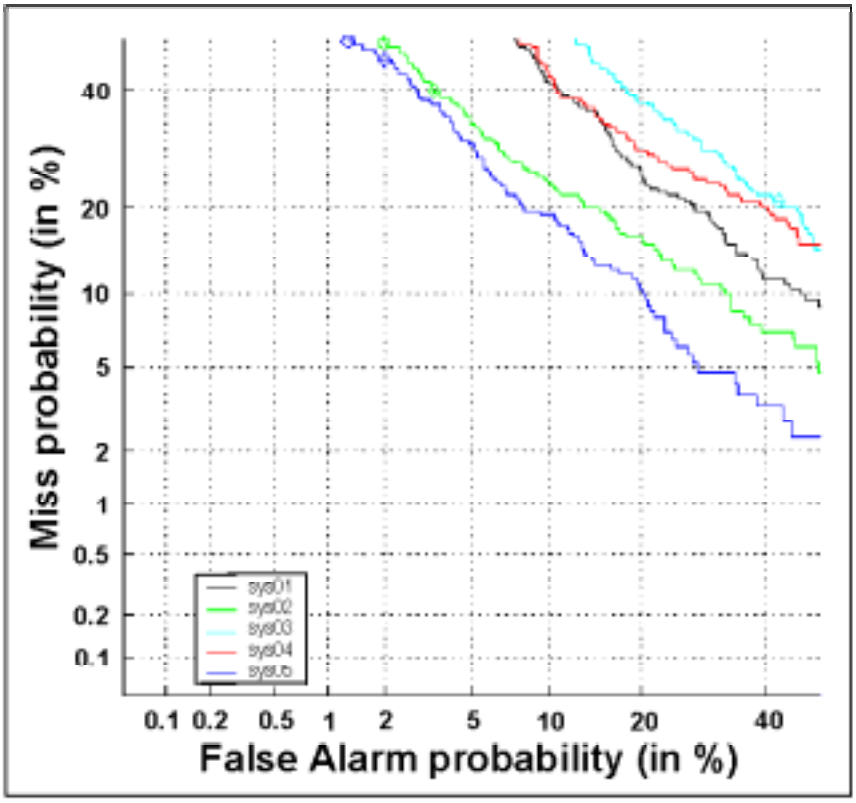

Miss probability (in %)
40
20
10
5
2
1
0.5
0.2
0.1
0.1 0.2 0.5 1 2 5 10 20 40
False Alarm probability (in %)
sys01
sys02
sys03
sys04
sys05

Figure 6

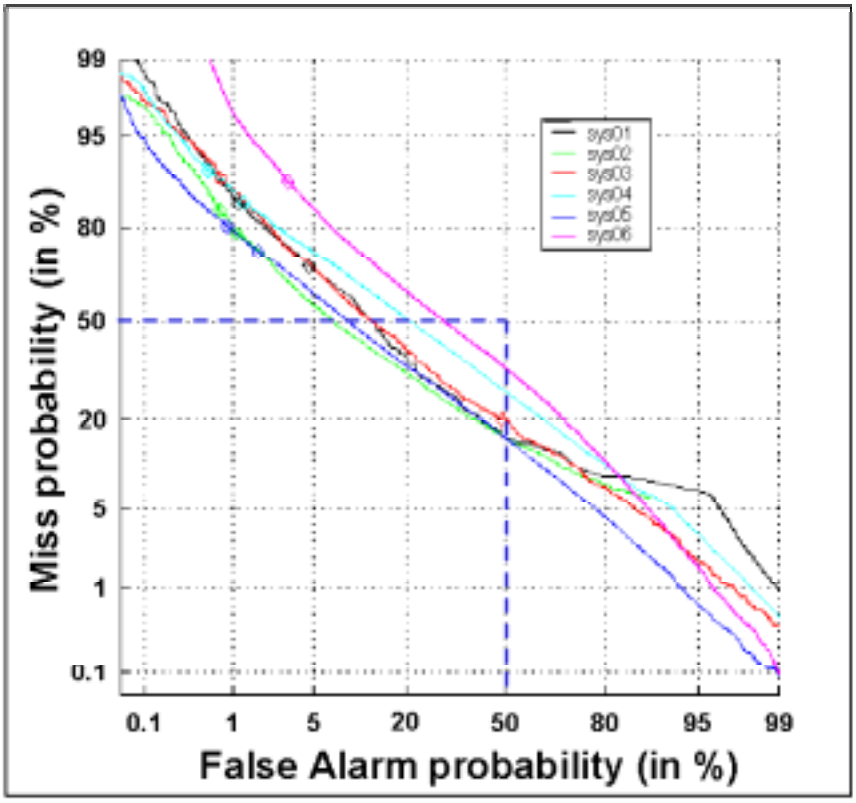

Miss probability (in %)
99
95
80
50
20
5
1
0.1
0.1
1
5
20
50
80
95
99
False Alarm probability (in %)
sys01
sys02
sys03
sys04
sys05
sys06

Figure 7

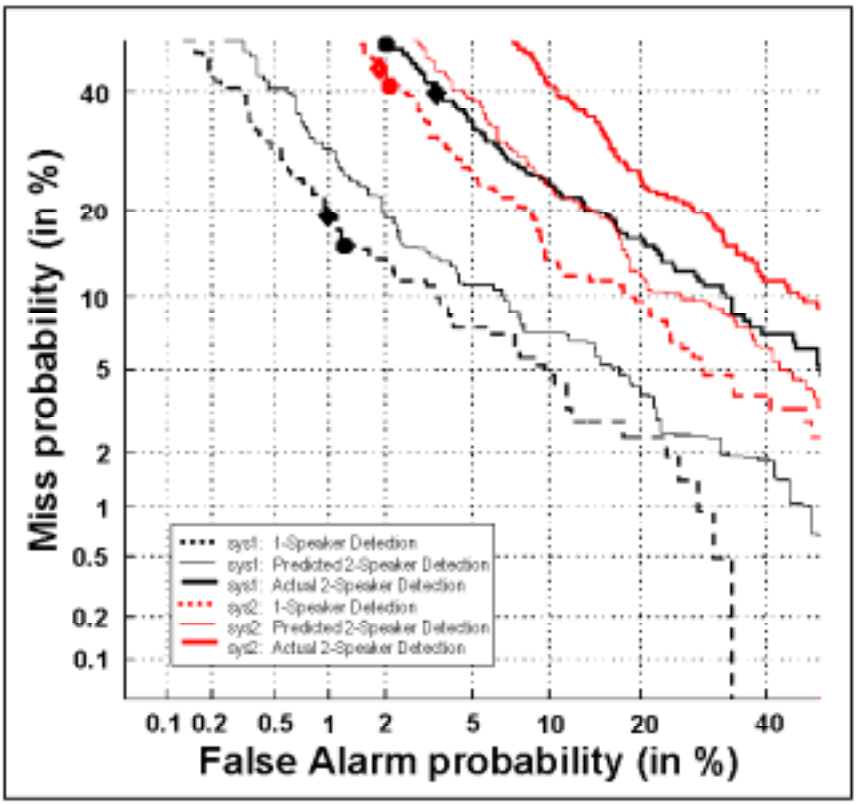

Miss probability (in %)
40
20
10
5
2
1
0.5
0.2
0.1
0.1 0.2 0.5 1 2 5 10 20 40
False Alarm probability (in %)
sys1: 1-Speaker Detection
sys1: Predicted 2-Speaker Detection
sys1: Actual 2-Speaker Detection
sys2: 1-Speaker Detection
sys2: Predicted 2-Speaker Detection
sys2: Actual 2-Speaker Detection

Figure 8


The 1999 Speaker Recognition Competition
Best Performance
One-Speaker Detection
Dragon Systems
Fred Weber
The 1999 Speaker Recognition Competition
Best Performance
Two-Speaker Detection
MIT Lincoln Laboratory
Douglas Reynolds
Bob Dunn
Tom Quatieri
The 1999 Speaker Recognition Competition
Best Performance
Speaker Tracking
MIT Lincoln Laboratory
Douglas Reynolds
Bob Dunn
Tom Quatieri

Figure 9

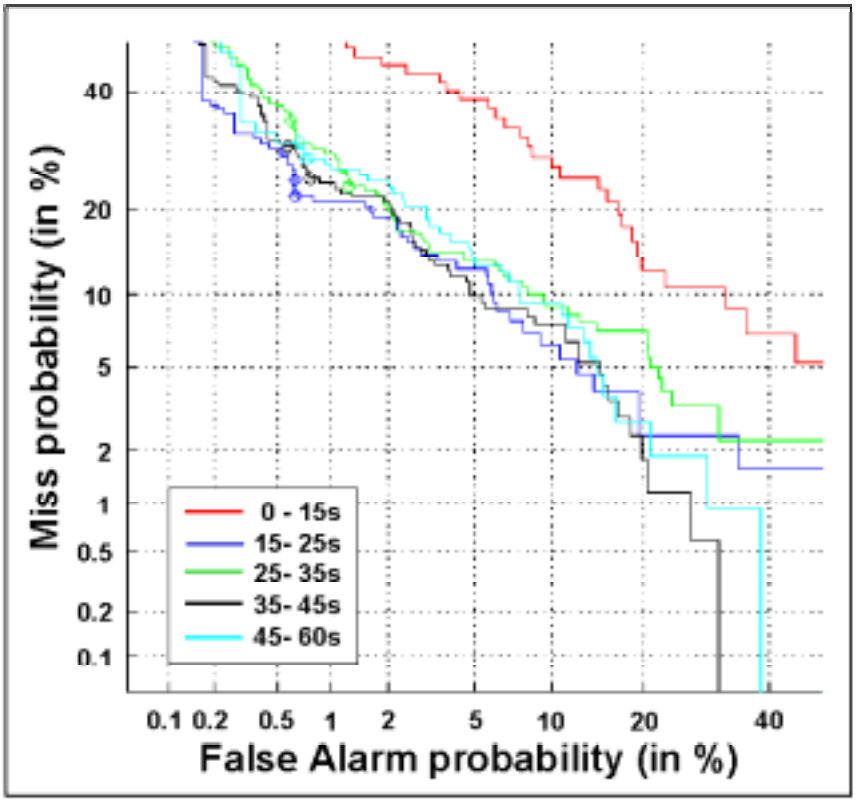

Miss probability (in %)
40
20
10
5
2
1
0.5
0.2
0.1
0.1 0.2 0.5 1 2 5 10 20 40
False Alarm probability (in %)
0 - 15s
15- 25s
25- 35s
35- 45s
45- 60s

Figure 10

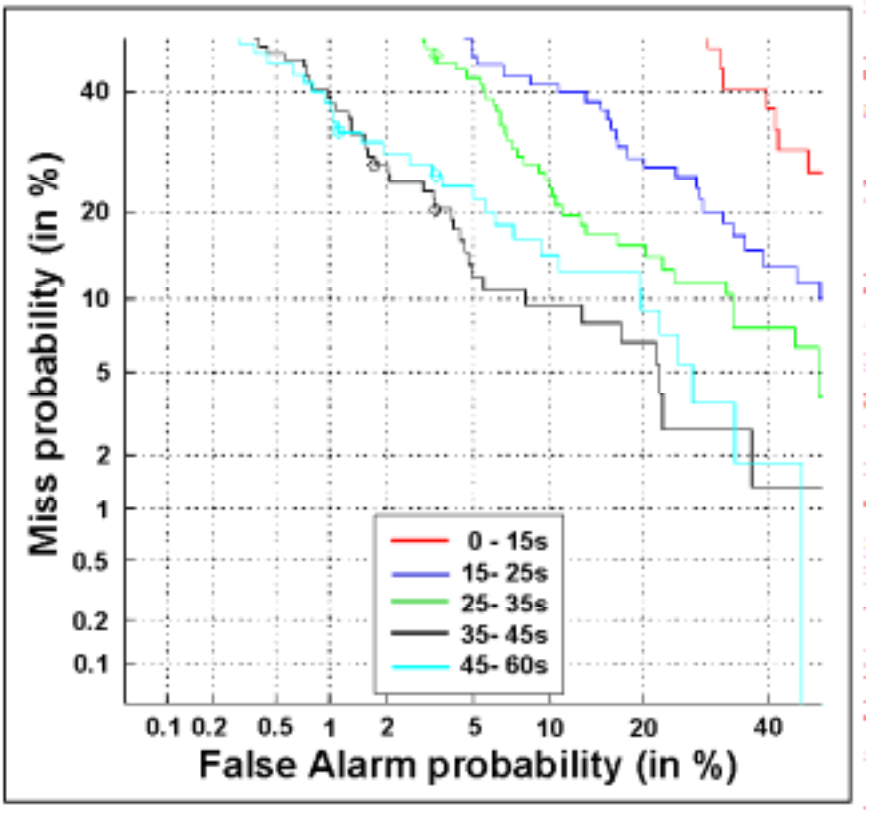

Miss probability (in %)
40
20
10
5
2
1
0.5
0.2
0.1
0.1 0.2 0.5 1 2 5 10 20 40
False Alarm probability (in %)
0 - 15s
15- 25s
25- 35s
35- 45s
45- 60s

Figure 11

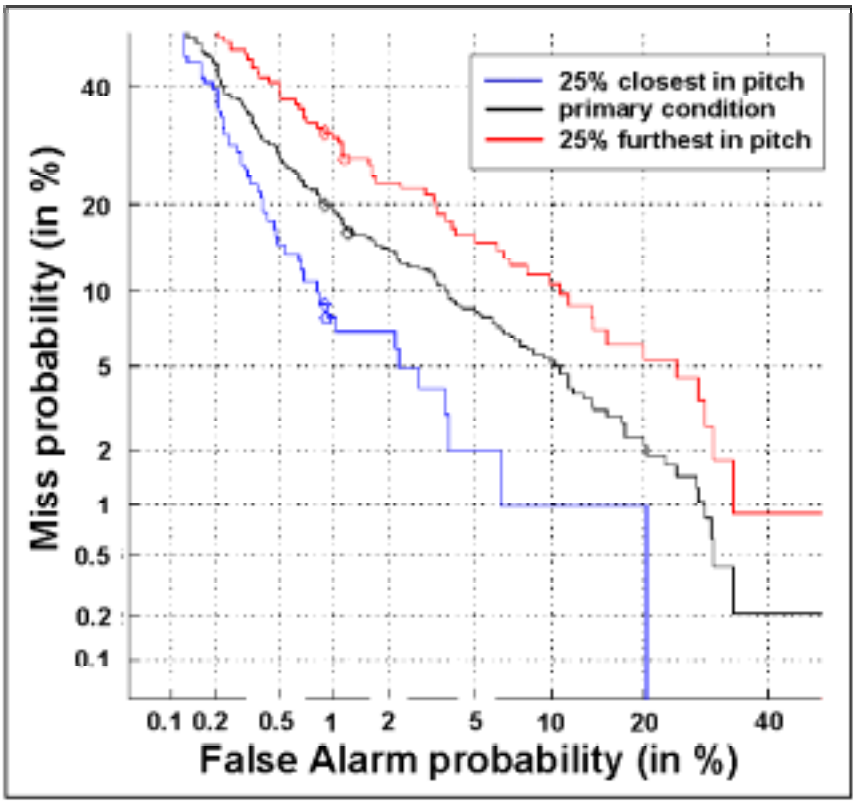

25% closest in pitch
primary condition
25% furthest in pitch
Miss probability (in %)
40
20
10
5
2
1
0.5
0.2
0.1
0.1 0.2 0.5 1 2 5 10 20 40
False Alarm probability (in %)

Figure 12

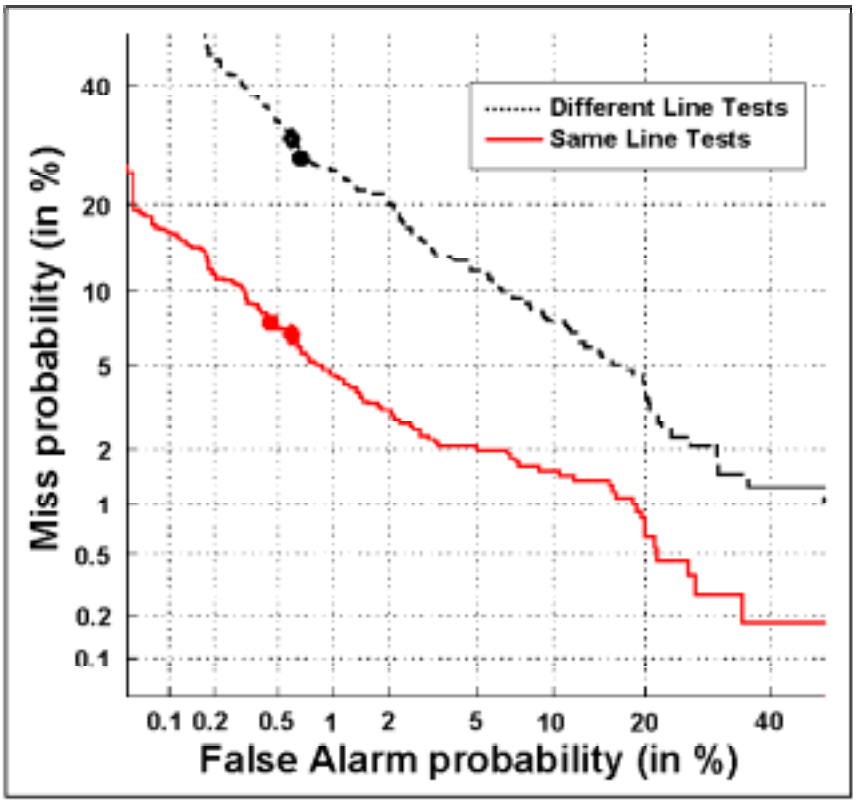

Different Line Tests
Same Line Tests
Miss probability (in %)
40
20
10
5
2
1
0.5
0.2
0.1
0.1 0.2 0.5 1 2 5 10 20 40
False Alarm probability (in %)

Figure 13

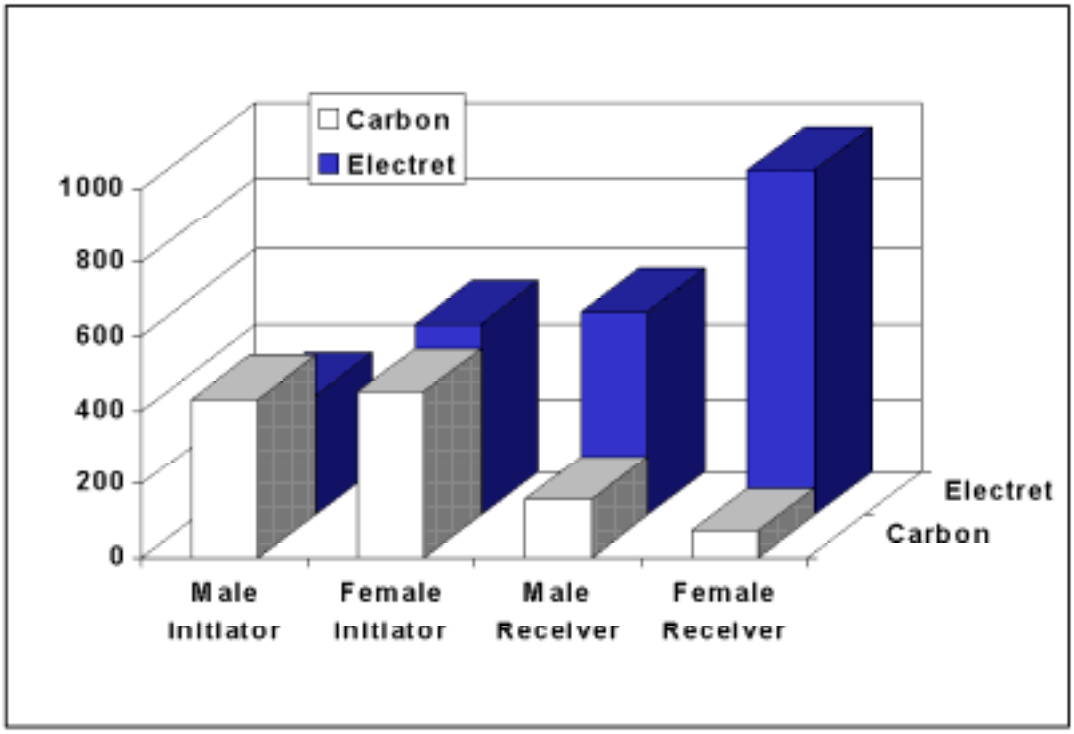
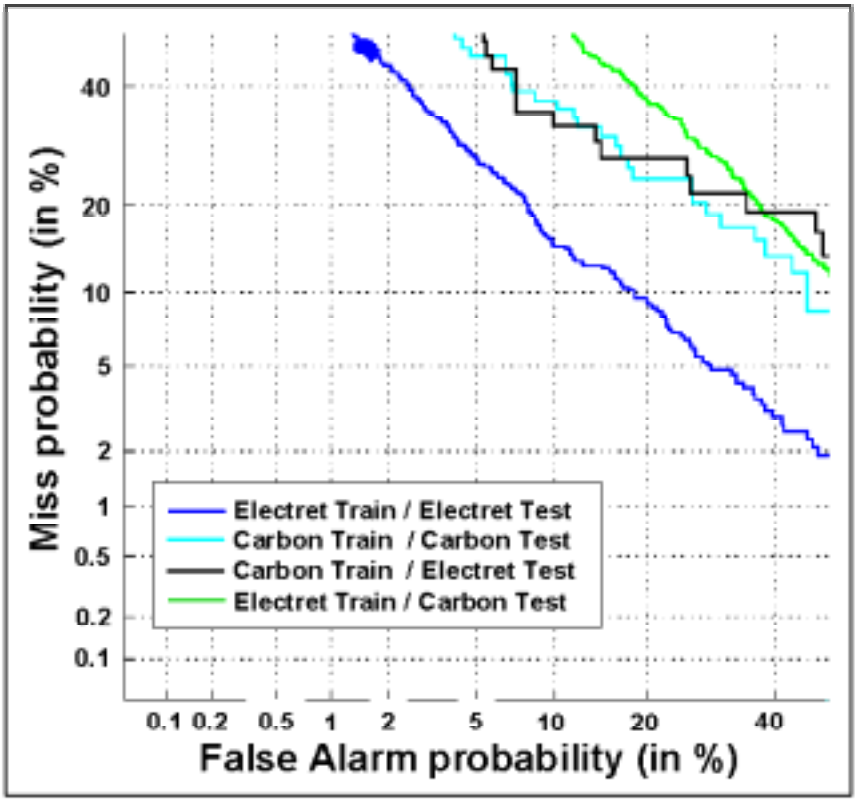

Carbon
Electret
1000
800
600
400
200
0
Male Initiator
Female Initiator
Male Receiver
Female Receiver
Electret
Carbon

Figure 14


Miss probability (in %)
40
20
10
5
2
1
0.5
0.2
0.1
Electret Train / Electret Test
Carbon Train / Carbon Test
Carbon Train / Electret Test
Electret Train / Carbon Test
0.1 0.2 0.5 1 2 5 10 20 40
False Alarm probability (in %)